Applied Statistics for Analytical Methods

Bachelor's Degree in Biotechnology — University of Teramo
Academic Year 2026/2027 · Prof. Federico Fanti

ANALYTICAL CHEMISTRY · 2ND YEAR

Lecture Outline

01

Measurement Uncertainty

Why every measurement has uncertainty and how to express it correctly

02

From Data to Knowledge

Types of data and analytical signal

03

Types of Error

Gross, systematic, and random errors

04

Precision, Trueness, Accuracy

VIM/ISO 5725 definitions and target diagrams

05

Descriptive Statistics

Mean, median, standard deviation, RSD

06

Normal Distribution and Inference

Gaussian curve, confidence interval, outlier tests

07

GUM Uncertainty and Calibration

Type A/B contributions, linear regression, residual analysis



Why Every Measurement Has an Uncertainty

It is impossible to perform a chemical analysis with results free of uncertainty. Every measurement is affected by errors of various kinds, and the task of the analytical chemist is to minimize errors and estimate their magnitude with acceptable accuracy.

Step 1

Establish the **maximum tolerable error** in the measurement before starting the analysis

Step 2

Perform **replicate measurements** to estimate variability and calculate the mean value

Step 3

Express the result as **value ± uncertainty** with correct units and significant figures

How to Correctly Express a Result

An analytical result is always expressed in the form:

$$x = \bar{x} \pm U \quad \text{[unit of measurement]}$$

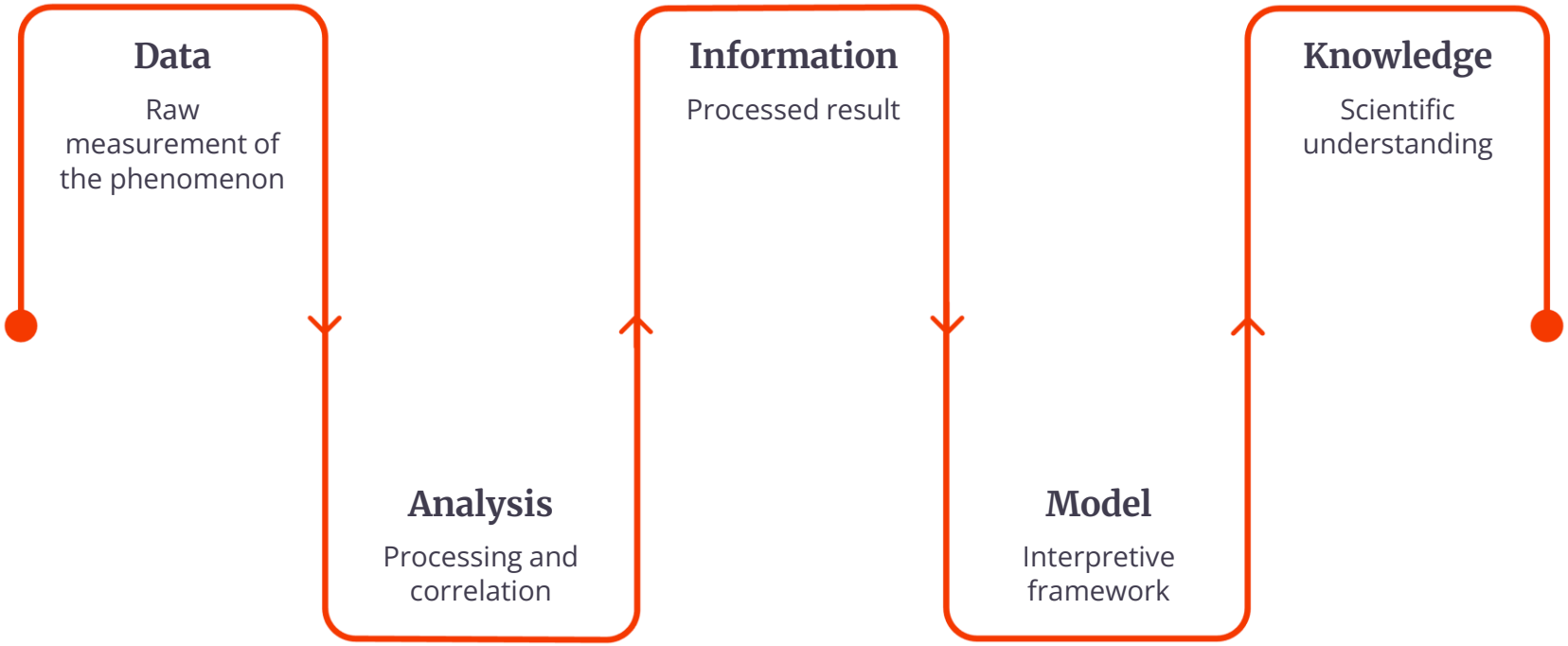
where **U** is the expanded uncertainty (see GUM section). The **significant figures** of the uncertainty determine the number of digits in the central value.

Example: blood glucose concentration = (5.42 ± 0.08) mmol/L

Practical rules

- Uncertainty is rounded to 1–2 significant figures
- The central value is rounded to the same decimal place as the uncertainty
- Always indicate the unit of measurement
- Specify the confidence level (e.g. k = 2, 95%)

From Data to Information to Knowledge



Data are elementary pieces of information that describe particular aspects of a phenomenon. On their own, data have no meaning: analysis is required to correlate them with significant aspects of the sample. For example, the chemical composition of blood only acquires meaning through a model of the human body and its pathologies.

Types of Data

Quantitative Data

Numerical value + unit of measurement
E.g.: *the temperature of water is 298.2 K*

Foundation of **hard sciences**: disciplines based on rigorous data connected by mathematical models.

Qualitative Data

Labels, descriptors, categories
E.g.: *"the water is warm"*

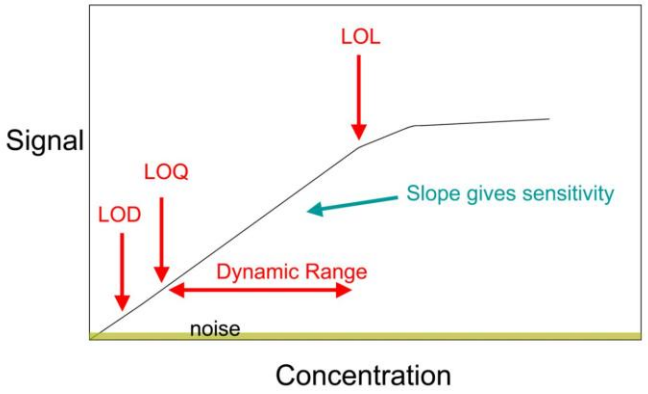
Difficult to standardize and reproduce (e.g. sensory analysis).

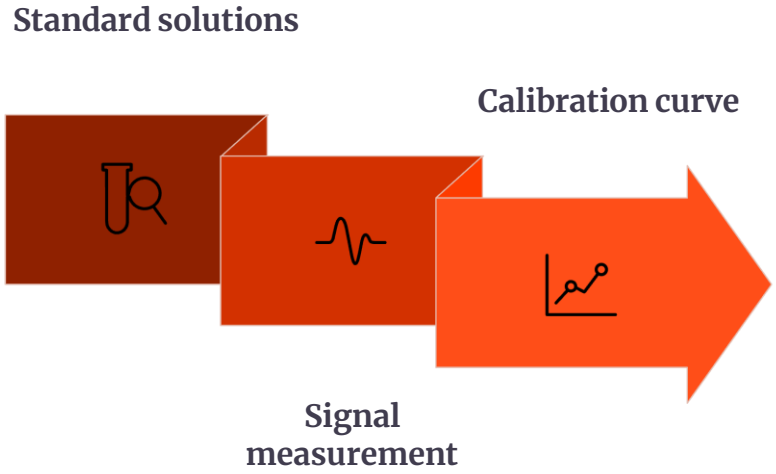
An individual's data includes: height, weight, skin color, concentration of chemical compounds in the blood, DNA composition, clothing and shoe size. Galilean science is founded on quantitative data.

The Analytical Signal and the Calibration Curve

The **analytical signal** is a quantity proportional to the amount of analyte present in the sample. The necessary condition is the existence of a stoichiometric (mathematical) relationship between the analyte and the signal source.

As the concentration of the analyte increases, the signal increases proportionally, allowing the construction of the **calibration curve**.





CHAPTER 1

Types of Error in Chemical Analysis

Gross Errors

Macroscopic errors that are easily identifiable (e.g. wrong sample, obvious contamination). They must be eliminated.

Systematic Errors

They follow defined laws and cause overestimation or underestimation of the true value. Identifiable and correctable.

Random Errors

Unpredictable fluctuations in experimental parameters. Cannot be eliminated, but can be minimized and quantified statistically.

The Concept of Error: Two Meanings

Meaning 1 — Error as a difference

The error represents the difference between the measured value and the "true" or "known" value:

$$E = x_i - x_t$$

Meaning 2 — Error as uncertainty

The error indicates the estimated uncertainty in a measurement or experiment, expressed as an interval around the central value.

The term **bias** indicates the systematic error of the population:
bias = μ − xt
where μ is the population mean and xt is the true value.

What can be done is: **minimize errors and estimate their magnitude with acceptable accuracy.**

Systematic Errors: Classification



Instrumental

Incorrect calibration, improper use of glassware and instruments, unsuitable instruments.



Method-Related

Non-ideal behavior of reagents and reactions, unsuitable experimental conditions (e.g. unexpected solubility, inadequate calcination times).



Personal

Distraction, ignorance of the procedure, physical impairments (astigmatism, color blindness), unconscious bias (e.g. stopping titration near the expected volume).

Systematic errors depend on causes that act according to **defined laws** and can be identified and corrected.

Systematic Errors: Constant and Proportional

Constant Errors

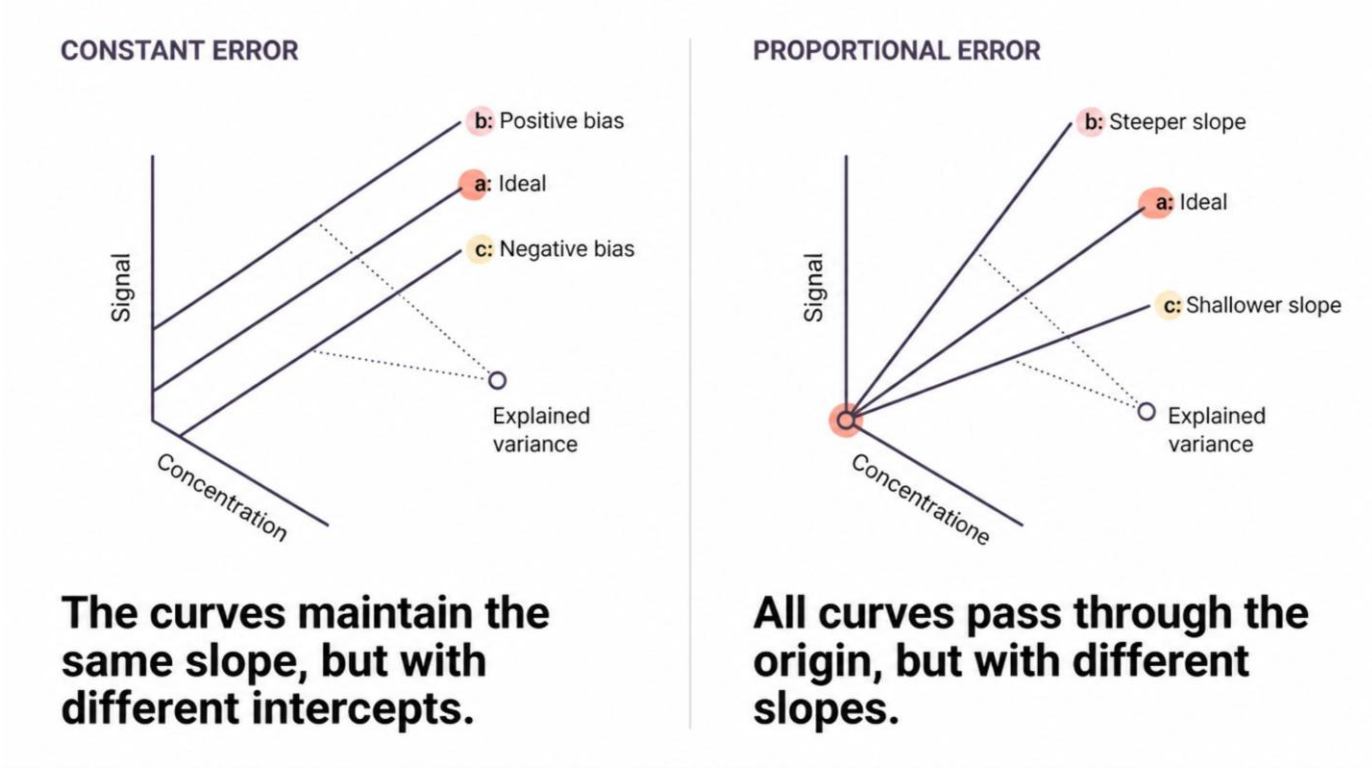
Shift the calibration curve parallel to the ideal line. The magnitude of the error is **independent of concentration**.

- Negative bias: losses due to solubility in gravimetry
- Positive bias: extraneous absorption in spectrophotometry

Proportional Errors

Modify the slope of the calibration curve. The magnitude of the error is **proportional to concentration**.

- Negative bias: analyte losses due to inefficient extractions
- Positive bias: matrix effect amplifying the signal



How to Identify and Eliminate Systematic Errors

1 Certified Reference Materials (CRM)

Analysis of standard samples with a known value certified by accredited bodies.

2 Independent Method

Comparison with proven instrumentation or a reference method of established reliability.

3 Blank Analysis

Solution containing all components except the analyte: corrects for the contribution of the matrix and solvent.

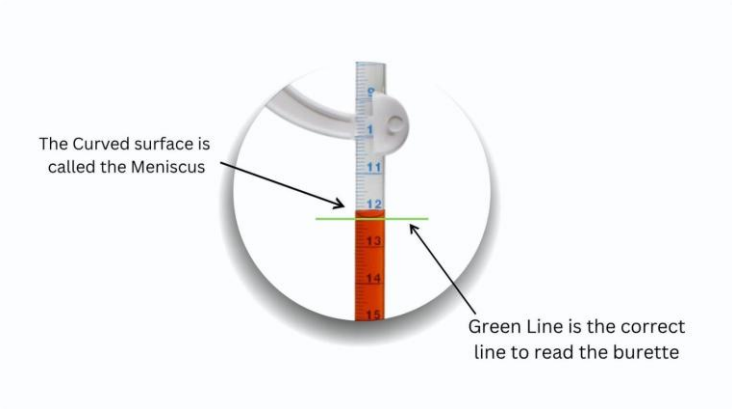
4 Recovery Tests (Spike)

Addition of known amounts of analyte to the real sample to verify the percentage recovery.

Random Errors: Characteristics and Example

Random errors (or indeterminate, *random*) cause a symmetric dispersion of data around the mean value. They are linked to indefinite fluctuations in: temperature, pH, pressure, humidity, ionic strength, and tolerances of balances and glassware.

Example — burette reading:
Burette with sensitivity = 0.1 mL. Reading: 9.68 mL. The estimate at the 0.01 mL level (1/10 of the sensitivity) is subject to random error.



Random errors **cannot be eliminated**, but can be reduced by operating carefully and quantified through statistical analysis of repeated measurements.

CHAPTER 2

Precision, Trueness, and Accuracy

The definitions follow the **VIM** (International Vocabulary of Metrology) and **ISO 5725** standards. Accuracy is the broadest concept and encompasses both trueness and precision.

Precision

The degree of agreement between results of successive measurements under the same conditions. Quantified by the standard deviation.

Trueness

The degree of agreement between the mean value of the results and the true value xt. Absence of systematic errors.

Accuracy

Encompasses both trueness and precision. An accurate measurement is both true and precise.

The Four Cases: Illustration with Targets



Neither accurate nor precise

Both systematic and random errors present. Points scattered far from the center.



Precise but not accurate

Only systematic errors. Points clustered but far from the center.



Accurate but not precise

Only random errors. Points scattered but centered on the true value.



Accurate and precise

Ideal goal. Points clustered and centered on the true value.

Repeatability and Reproducibility

Repeatability

Precision obtained under the **same conditions**: same operator, same instrument, same laboratory, short time interval. Minimum expected variability.

Reproducibility

Precision obtained under **different conditions**: different operators, instruments, or laboratories. Greater variability, but essential for interlaboratory validation.

ISO 5725 clearly distinguishes the two concepts. Reproducibility is always ≥ repeatability in terms of variance. Both are quantified using the standard deviation under their respective conditions.

CHAPTER 3

Descriptive Statistics: Sample and Population

Sample

Set of measurements under examination (finite). Notation: $\bar{x}$ for the mean, **s** for the standard deviation.

Population

Set of all possible measurements (theoretically infinite). Notation: **μ** for the mean, **σ** for the standard deviation.

The sample distribution can be represented in a **discrete** form (histogram). The population distribution can be represented by a **continuous curve** (normal or Gaussian curve). Sample statistics are estimates of the population parameters.

Mean and Median

Arithmetic Mean

Sum of all measurements divided by the number of measurements:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Sensitive to outliers.

Median

The middle value after sorting in ascending order. For even N: average of the two central values. Robust to outliers.

Example with an outlier:

$$X = \{12, 15, 20, 23, 24, 27, \mathbf{200}\}$$

Median = **23** · Mean = **45.9**

⚠ The outlier 200 shifts the mean to 45.9, far from the main cluster. The median (23) is much more representative in this case.

Standard Deviation

The **standard deviation** (or root mean square deviation) measures the dispersion of data around the expected value. It has the same unit of measurement as the observed values.

For the Population (σ)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

Used when the entire population is known.

For the Sample (s)

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

n−1 (degrees of freedom) is used when working with a sample of the population.

Relative Standard Deviation (RSD) and Coefficient of Variation (CV%)

$$\text{RSD} = \frac{s}{\bar{x}} \qquad \text{CV\%} = \frac{s}{\bar{x}} \times 100$$

Numerical Exercise — Descriptive Statistics

Determination of Fe(III) concentration in a water sample (mg/L), 7 replicates:

Replicate	1	2	3	4	5	6	7
Conc. (mg/L)	2.41	2.38	2.45	2.43	2.39	2.44	2.42

Calculations

$$\bar{x} = \frac{2{,}41 + 2{,}38 + 2{,}45 + 2{,}43 + 2{,}39 + 2{,}44 + 2{,}42}{7} = 2{,}417 \text{ mg/L}$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{6}} = 0{,}024 \text{ mg/L}$$

Results

2.417

mg/L — Mean

0.024

mg/L — Std. Dev.

1.0%

CV% — Precision

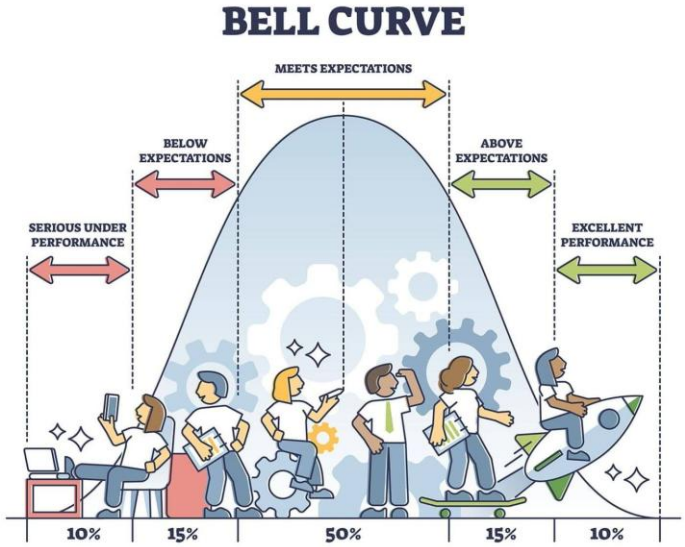
CHAPTER 4

The Normal (Gaussian) Distribution

The **Gaussian function** describes the distribution of random errors:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty$$

where **μ** is the population mean and **σ** the standard deviation. The total area under the curve equals **1** (representing the total probability).



μ ± 1σ

68.3% of measurements

μ ± 2σ

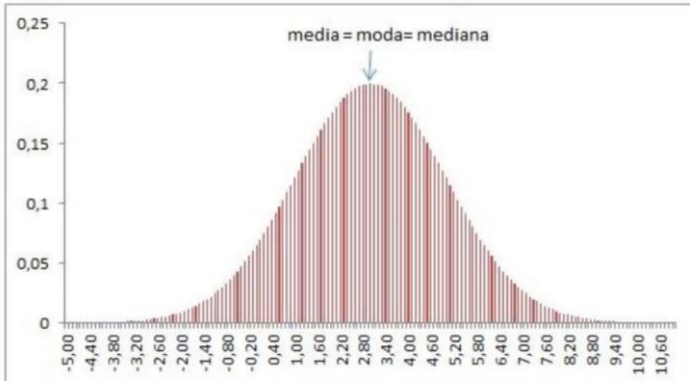
95.4% of measurements

μ ± 3σ

99.7% of measurements

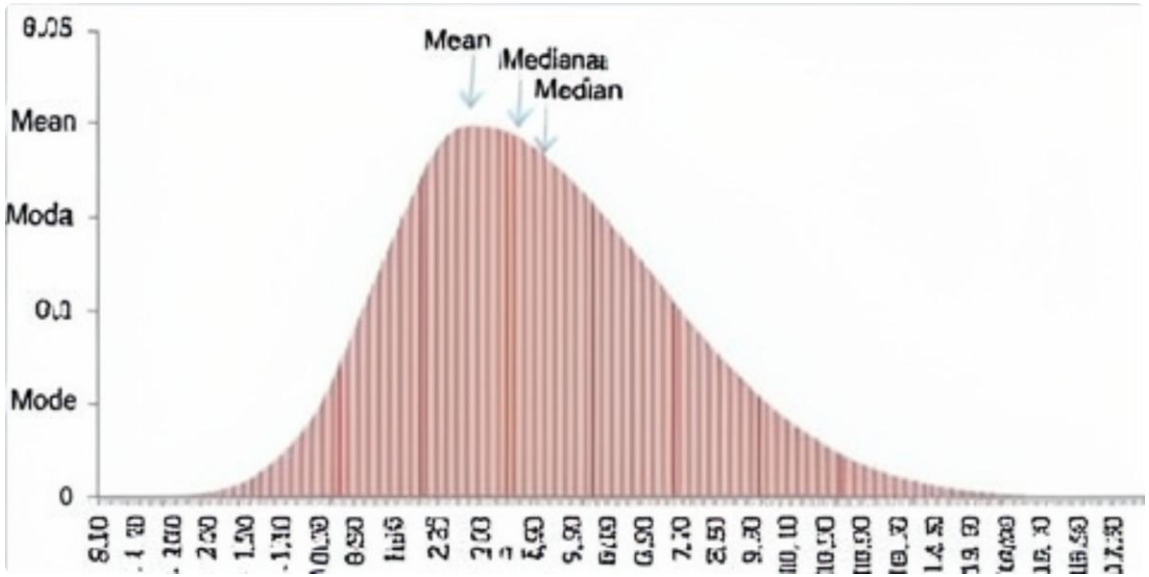
Distributions: Symmetric

When data does not follow a normal distribution, the relative position of **mean, median and mode** changes in a characteristic way.



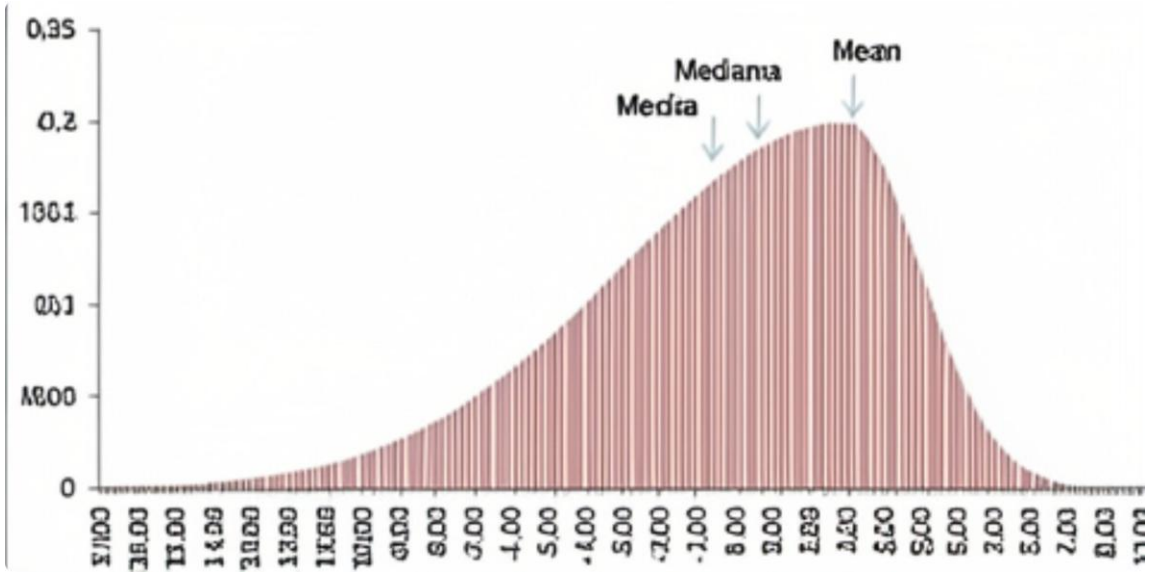
Symmetric Distribution

Mean = Median = Mode. Typical of pure random errors.



Positive Skewness (right tail)

Mean > Median > Mode. The tail extends to the right.



Distributions: Negative Skewness (left tail)

Mode > Median > Mean. The tail extends to the left.

Basic Inference: Standard Error and t-Distribution

Standard Error of the Mean

The sample mean is itself a random variable. Its standard deviation is the **standard error**:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

As n increases, the standard error decreases: more measurements → more reliable estimate.

Student's t-Distribution

For small samples ($n < 30$) with unknown σ, the **Student's t-distribution** is used with $\nu = n-1$ degrees of freedom.

Confidence Interval

$$\bar{x} \pm t_{\nu,\alpha} \cdot \frac{s}{\sqrt{n}}$$

The t-value depends on the confidence level (e.g. 95%) and the degrees of freedom.

Tests for Outliers: Dixon Q Test and Grubbs Test

Dixon Q Test

Suitable for small samples ($3 \le n \le 10$). The ratio between the difference of the suspect value from its neighbor and the total range is calculated:

$$Q = \frac{| x_{sospetto} - x_{vicino} |}{x_{max} - x_{min}}$$

If Q > Qcritical (from table), the value is an outlier at the chosen confidence level.

Grubbs Test

Suitable for larger samples. The G statistic is calculated:

$$G = \frac{| x_{sospetto} - \bar{x} |}{s}$$

If G > Gcritical, the value is statistically anomalous. More powerful than the Q test for n > 10.

An outlier should only be removed if a real experimental cause is identified (e.g., transcription error, contamination). It should never be discarded simply because it is "far from the others."

CHAPTER 5

Measurement Uncertainty according to GUM

The **GUM** (Guide to the Expression of Uncertainty in Measurement, JCGM 100:2008) provides the international framework for the evaluation and expression of uncertainty. It distinguishes two categories of contributions.

Type A Contribution

Evaluated by statistical analysis of a series of observations (e.g. repeatability uncertainty). Calculated as the standard error of the mean: uA = s/√n.

Type B Contribution

Evaluated by methods other than statistical analysis: calibration certificates, manufacturer specifications, literature data, expert judgment.

Sources of Uncertainty and Expanded Uncertainty

Repeatability (Type A)

uA = s/√n from the series of replicates

Instrumental Calibration (Type B)

From the instrument calibration certificate

Weighing (Type B)

From the resolution and calibration of the balance

Final Volume (Type B)

From the tolerance of volumetric glassware

Internal Standard and Calibration Solution (Type B)

From the purity and preparation of the standards

Combined and Expanded Uncertainty

$$u_c = \sqrt{u_A^2 + u_{B1}^2 + u_{B2}^2 + \ldots} \qquad U = k \cdot u_c \quad (k = 2,\ \text{confidence} \approx 95\%)$$

CHAPTER 6

Calibration and Linear Regression

Linear regression is the fundamental statistical tool for building the calibration curve. It allows predicting the values of a dependent variable (signal Y) from an independent variable (concentration X).

Simple Regression

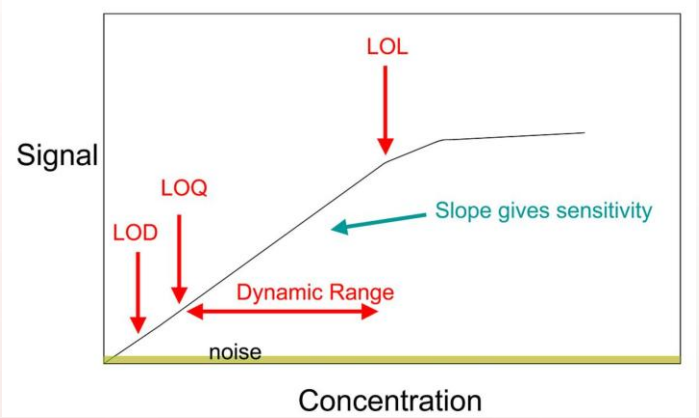
A single explanatory variable $X \rightarrow Y = \beta_0 + \beta_1 X + \varepsilon$

Multiple Regression

Several explanatory variables $X_1, X_2, \ldots, X_p \rightarrow$ more complex models

Correlation

Study of the association between quantitative variables without causal prediction



The Simple Linear Regression Model

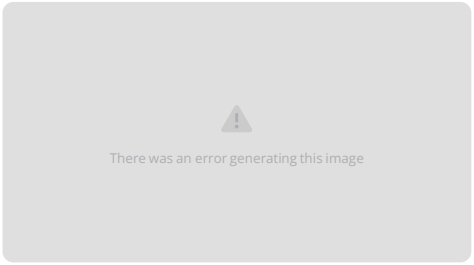
The simple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where:

- β_0 = population intercept (value of Y when X = 0)
- β_1 = population slope: change in Y per unit of X
- ε_i = random error for the i-th observation

The sign of β_1 indicates whether the relationship is positive or negative. In calibration, β_1 represents the **analytical sensitivity**.



Least Squares Method

The sample estimates b_0 and b_1 are obtained by minimizing the **sum of squared errors** (SSE). The formulas are:

Slope Estimate

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - X)(Y_i - Y)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Intercept Estimate

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Sample Regression Line Equation

$$\hat{Y}_i = b_0 + b_1 X_i$$

where $\hat{\mathbf{Y}}_i$ is the predicted value of Y for observation i. The line always passes through the point ($\bar{X}$., $\bar{Y}$

Interpolation, Extrapolation and Inverse Prediction

Interpolation

The prediction of Y corresponds to a value of X **within** the interval used to estimate the model. Reliable and recommended.

Extrapolation

The prediction of Y corresponds to a value of X **outside** the interval. Risky: linearity may not hold beyond the calibration range.

Inverse Prediction (Calibration)

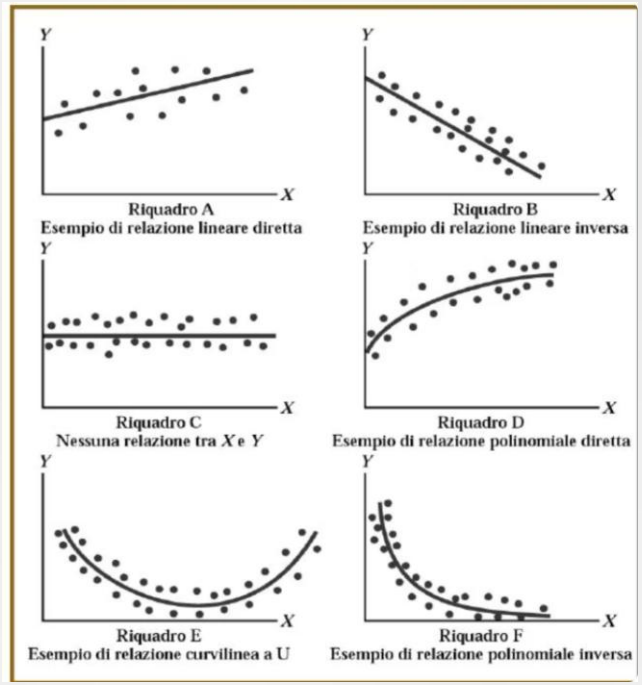
Given an instrumental response y_0 measured on an unknown sample, the concentration is derived as:

$$x_0 = \frac{y_0 - b_0}{b_1}$$

This is the fundamental operation in quantitative analysis.

Types of Relationships in the Scatter Plot

The choice of the appropriate mathematical model is suggested by the distribution of points in the scatter plot: direct linear relationship (A), inverse linear relationship (B), no relationship (C), direct polynomial relationship (D), U-shaped curvilinear relationship (E), inverse polynomial relationship (F).



Decomposition of Variability: SST = SSR + SSE

$$SST = SSR + SSE$$

SST — Total Variability

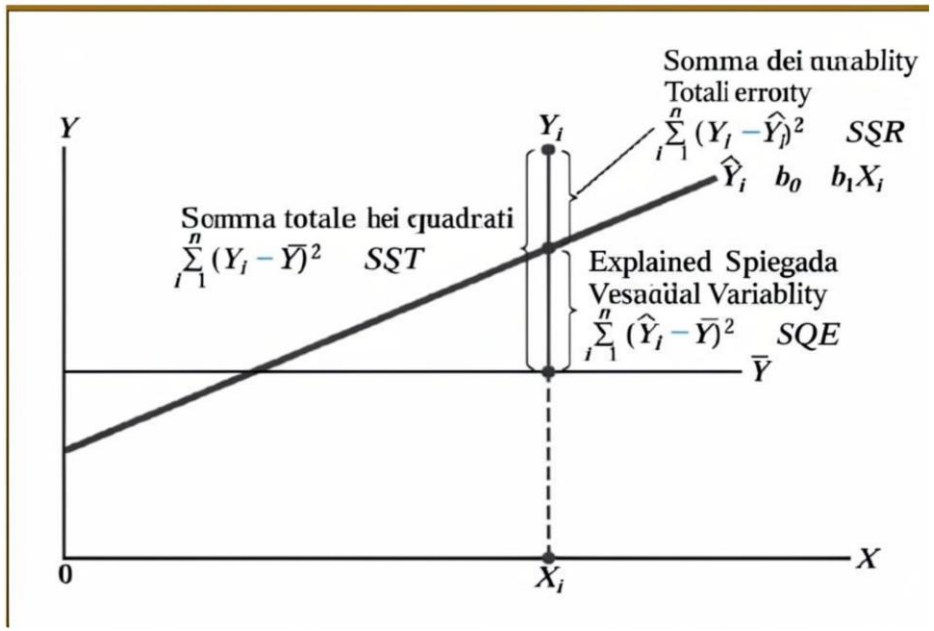
$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

SSR — Explained Variability

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

SSE — Unexplained Variability

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$



Coefficient of Determination r^2 and Pearson Correlation r

Coefficient of Determination r^2

$$r^2 = \frac{SSR}{SST}$$

Measures the **proportion of variability in Y explained by X**. Ranges between 0 and 1. An $r^2 = 0.999$ indicates that 99.9% of the variability is explained by the model.

Pearson Correlation Coefficient r

$$r = \frac{\sum(X_i - X)(Y_i - Y)}{\sqrt{\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2}}$$

Ranges between −1 and +1. The sign indicates the direction of the relationship.

⚠ **Warning:** a high r^2 alone does not prove the linearity of the model. It is essential to always check the residuals plot to rule out systematic patterns.

Numerical Exercise — Calibration Curve

Spectrophotometric calibration for Fe(III) at 5 concentration levels:

Level	Conc. X (mg/L)	Absorbance Y	$\hat{Y}$ (predicted)	Residual e_i
1	0.50	0.062	0.061	+0.001
2	1.00	0.121	0.120	+0.001
3	2.00	0.238	0.239	−0.001
4	4.00	0.479	0.477	+0.002
5	6.00	0.714	0.715	−0.001

$$b_1 = 0{,}119 \text{ L/mg} \quad b_0 = 0{,}002$$

$$r^2 = 0{,}9999 \quad \Rightarrow \text{excellent linearity}$$

For an unknown sample with $y_0 = 0.350$:

$$x_0 = \frac{0{,}350 - 0{,}002}{0{,}119} = 2{,}92 \text{ mg/L}$$

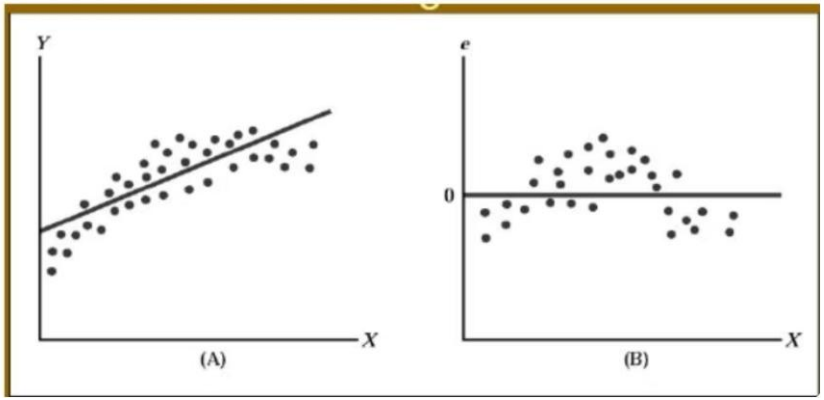
CHAPTER 7

Residual Analysis: Model Verification

The **residual** is the estimate of the error made in predicting Y_i through $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i$$

The residual plot (e_i) vs X is the fundamental diagnostic tool. If the model is adequate, the residuals must be distributed randomly around zero, without systematic patterns.



If a curvilinear structure is evident in the residuals, the linear model is not adequate: a polynomial or non-linear model should be considered.

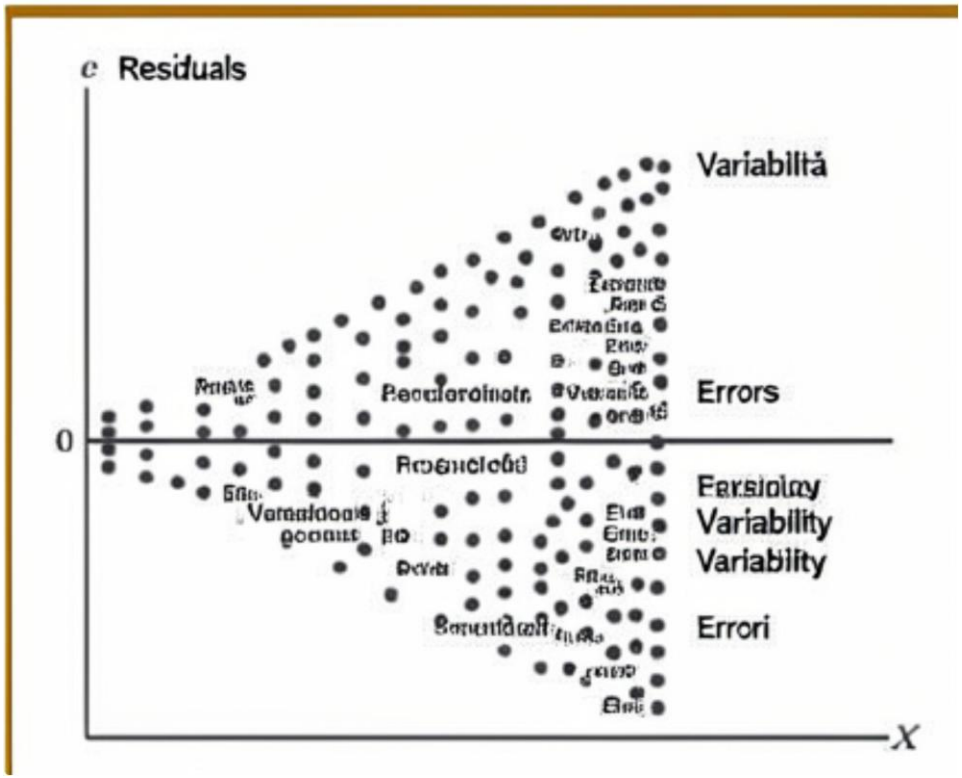
Homoscedasticity and Heteroscedasticity

Homoscedasticity

The variance of the errors is **constant** across values of X. This is a fundamental assumption of ordinary least squares (OLS) regression. The residual plot shows a uniform band around zero.

Heteroscedasticity

The variance of the errors **increases** (or varies) with X. Common in analytical chemistry over wide concentration ranges. The residual plot shows a "funnel-shaped" pattern.



In the presence of heteroscedasticity, **weighted least squares** regression is used, with weights proportional to $1/x$ or $1/x^2$, to reduce the influence of points with higher variance. This topic is discussed in depth in the second Teaching Unit on method validation.

Effect of Systematic Errors on the Calibration Curve

Systematic errors modify the calibration line in a predictable way, affecting the estimation of the unknown concentration x_0.

Positive Constant Error

Shifts the line upward (higher intercept). The unknown concentration is **underestimated** if the blank is not corrected.

Negative Constant Error

Shifts the line downward (lower intercept). The unknown concentration is **overestimated**.

Positive Proportional Error

Increases the slope (apparently higher sensitivity). The unknown concentration is **underestimated**.

Negative Proportional Error

Decreases the slope. The unknown concentration is **overestimated**. Typical of inefficient extractions.



Connection to Method Validation

The statistical concepts developed in this lesson form the foundation for **analytical method validation**, covered in the second Teaching Unit. The main validation parameters are:

Linearity and Range

Verification of r^2, residual analysis, lack of fit test

LOD and LOQ

Limit of detection and quantification, derived from the blank's s and the slope

Recovery and Selectivity

Spike tests, interferents, matrix effect: all evaluated with statistical tools

Measurement Uncertainty

Complete budget according to GUM, with all Type A and Type B contributions identified

Summary of Key Concepts

Uncertainty

Every measurement has uncertainty. Expressed as value ± U with correct units and significant figures.

Errors

Systematic (bias, correctable) and random (reducible, quantifiable). Bias = μ − xt.

Precision/Accuracy

Precision = dispersion (s, CV%). Accuracy = absence of bias. Trueness = both combined.

Gaussian

68–95–99.7% rule. Total area = 1. Confidence interval: $\bar{x} \pm t \cdot s/\sqrt{n}$.

GUM

Type A (statistical) and Type B (other) contributions. Expanded uncertainty U = k·uc with k = 2.

Regression

$\hat{Y} = b_0 + b_1X$. Least squares. r^2 and residual analysis. Inverse prediction: $x_0 = (y_0 - b_0)/b_1$.

Statistics is not optional: it is an integral part of the scientific method in analytical chemistry. Any result without uncertainty is incomplete.